DOCUMENT RESUME

ED 475 839 TM 034 865

AUTHOR	Dimitrov, Dimiter M.
TITLE	Reliability of True Cutting Scores for Rasch Calibrated Items.
PUB DATE	2003-04-00
NOTE	26p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Cutting Scores; Item Response Theory; *Reliability; *True Scores
IDENTIFIERS	*Calibration; *Rasch Model

ABSTRACT

This paper provides formulas for expected true-score measures and reliability of binary items as a function of their Rasch difficulty parameters when the trait distribution is normal or logistic. With the proposed formula, one can evaluate the theoretical values of classical reliability indexes for norm-referenced and criterion-referenced interpretations without information about raw-score or trait scores of persons from the target population. This is achieved by representing the theoretical (marginalized) values of the true-score components of reliability indexes as functions of the item difficulty parameter. As the analytic forms of such functions are developed for individual items (and then "summarized" at test level), one can know the population values of true-score measures and reliability for a set of Rasch calibrated binary items prior to their administration. An example for the application of the proposed formulas and their empirical validation is also provided. (Contains 2 tables, 2 figures, and 31 references.) (Author)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 475 839

Reliability of True Cutting Scores for Rasch Calibrated Items

Dimiter M. Dimitrov
Kent State University
(e-mail:ddimitro@kent.edu)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
D. Dimitrov
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM034865

Paper Presented at the Annual Meeting of the
American Educational Research Association
Chicago, IL, April 21-25, 2003

BEST COPY AVAILABLE

Abstract

This papere provides formulas for expected true-score measures and reliability of binary items as a function of their Rasch difficulty parameters when the trait distribution is normal or logistic. With the proposed formulas, one can evaluate the theoretical values of classical reliability indexes for norm-referenced and criterion-referenced interpretations without information about raw-score or trait scores of persons from the target population. This is achieved by representing the theoretical (marginalized) values of the true-score components of reliability indexes as functions of the item difficulty parameter. As the analytic forms of such functions are developed for individual items (and then "summarized" at test level), one can know the population values of true-score measures and reliability for a set of Rasch calibrated binary items *prior* to their administration. An example for the application of the proposed formulas and their empirical validation is also provided.

Reliability of True Cutting Scores for Rasch Calibrated Items

Despite disadvantages of the True-Score Model (TSM) in metric development and accuracy of measurement compared to item-response theory models (e.g., Hambleton and Jones, 1993) and Rasch models (e.g., Linacre, 1997; Smith, 2000, 2001), the TSM has been and is still used in test development and test score analysis. Traditionally, true-score measures represent a common focal point between test developers and practitioners as they place the scores and their accuracy in the original scale of measurement [e.g., number-right (*NR*) scores.], True scores are are readily interpretable and, for example, when pass-fail decisions are made, a cutting score is typically set on the domain-score scale (e.g., Hambleton, Swaminathan, & Rogers, 1991, p. 85). Recent debates and editorial policies on issues of reliability (e.g., Dimitrov, 2002; Sawilowsky, 2000; Thompson & Vacha-Haase, 2000) also indicate the necessity of adequate understanding and estimation of TSM reliability and standard error of measurement at sample and population level. In this paper, marginal population values of true-score measures for individual binary items are determined from their Rasch difficulty parameters. In a previous work (Dimitrov, 2002) this has been achieved only for the marginal *item score* and *item error variance*. This paper completes the work by providing analytic evaluations for *item true variance*, *item reliability*, *classical test reliability*, and *dependability of cutting scores* for criterion-referenced ("pass/fail") decisions. The proposed formulas have theoretical value and can be very useful in test development, score analysis, and simulation studies. For example, given a bank of binary items calibrated with the dichotomous Rasch model (Rasch, 1960), one can select items with known true-score measures and reliability *prior* to administering the test.

It is important to note that the information provided with the proposed formulas and the information obtained through Rasch analysis can complement (not replace or exclude) each other in measurement analysis. For example, the TSM reliability evaluated with the method developed in this article provides more information about the accuracy of measurement at population level relative to classical coefficients such as Cronbach's alpha (Cronbach, 1951), but it cannot replace

the information provided by Rasch reliability measures for locating persons on the underlying trait (e.g., Linacre, 1996, 1997; Wright, 2001). Other comments on this issue that follow later in the text also support the argument that researchers and practitioners will benefit from combining Rasch measurement information with TSM information about population measures provided by formulas developed in this paper.

Theoretical Framework

As the title of the paper indicates, the reliability of true cutting scores for Rasch calibrated binary items is a primary target of the study. It should be noted, however, that the "intermediate" results (formulas) developed in this paper in achieving this task have their own (methodological and technical) value in Rasch test development and analysis. Therefore, the paper is organized by (1) presenting true-score measures and reliability for individual binary items as a function of their Rasch difficulty parameter, (2) "summarizing" the results for individual items to obtain evaluations for true-score measures at test level (e.g., error variance and true variance for the *NR* score), and (3) using the true-score measures at test level to evaluate the theoretical reliability for both norm-referenced and criterion-referenced reliability.

With the dichotomous Rasch model, the probability for correct answer on item *i* with difficulty δ_i for a person with a trait score θ is

$$P_i(\theta) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)}. \qquad (1)$$

As $P_i(\theta)$ is also the true score on item *i* for a person at θ, the *item score* (marginal probability for correct response on the item) is

$$\pi_i = \int_{-\infty}^{\infty} P_i(\theta)\varphi(\theta)d\theta, \qquad (2)$$

where $\varphi(\theta)$ is the probability density function (*pdf*) for the population trait distribution. With this, the *marginal NR score* for a test of *n* binary items is then

$$\mu = \sum_{i=1}^{n} \pi_i \qquad (3)$$

and the *expected domain score* is $\pi = \mu/n$ (in terms of percentages: $\pi = 100\mu/n$).

Also, $P_i(\theta)[1 - P_i(\theta)]$ is the error variance for a binary item *i* at θ (Lord, 1980, p. 52). Therefore, the (marginal) *item error variance* is

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} P_i(\theta)[1 - P_i(\theta)]\varphi(\theta)d\theta. \qquad (4)$$

The (marginal) *error variance for the NR score* on a test of *n* dichotomous items is then

$$\sigma_e^2 = \sum_{i=1}^{n} \sigma^2(e_i). \qquad (5)$$

It is important to emphasize that σ_e^2 represents the accuracy of number-right scores and is not to be confused with the mean square measurement error (MSE_p) that represents the accuracy of trait scores on the logit scale with Rasch measurement models (e.g., Smith, 2001). Also, while the MSE_p is a sample statistic that requires information about the person's trait score, θ, σ_e^2 does not require such information because it is obtained through integration over the trait interval.

Closed form integral evaluations for π_i and $\sigma^2(e_i)$ in Equations 2 and 4, respectively, are provided in the next section. The population distribution for the underlying trait is assumed to be normal or logistic. The *pdf* of a logistic distribution (e.g., Evans, Hastings, & Peacock, 1993, p. 98) with the location at the origin of the scale is

$$\varphi(\theta) = \frac{\exp(\theta / c)}{c[1 + \exp(\theta / c)]^2}, \qquad (6)$$

where c is the scale parameter. This paper deals with two specific logistic distributions ($c = 1$ or $c = 1/2$) that yield exact integral evaluations for Equations 2 and 4 and capture normal-like ability shapes that may occur in practice with Rasch measurement (see Figure 1).

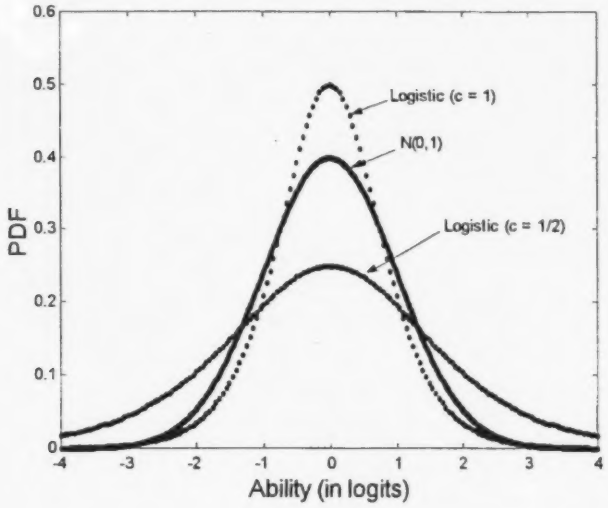

Figure 1. Probability density functions (PDF) of the standard normal distribution and two logistic distributions with scale parameters $c = 1$ and $c = ½$.

Formulas at Item Level

Item Score with the Normal Ability Distribution

With $P_i(\theta)$ for the dichotomous Rasch model and $\varphi(\theta)$ with $N(0,1)$, an exact closed form evaluation for the integral in Equation 2 does not exist. Therefore, an approximation formula was developed in two steps. First, using the computer program MATLAB (MathWorks, Inc., 1999), quadrature method evaluations were obtained for values of the Rasch item difficulty, δ_i in the interval from -6 to 6 with an increment of 0.01 on the logit scale. Second, the results were tabulated and then approximated with the four-parameter sigmoid function using the regression wizard of the computer program SigmaPlot 5.0 (SPSS Inc., 1998). The resulting approximation formula (with an absolute error smaller than 0.02) for the expected item mean is

$$\pi_i = -0.0114 + \frac{1.0228}{1 + \exp(\delta_i / 1.226)}. \quad (7)$$

Formula 7 can be used with any normal trait distribution, $N(\mu_\theta; \sigma_\theta)$, after transforming the item difficulty estimate: $\delta_i^* = (\delta_i - \mu_\theta)/\sigma_\theta$ (e.g., Smith, 2000). For n binary items, the expected number right-score, μ, is obtained with Equation 3; (the expected domain score is $\pi = \mu/n$).

Item Score with Logistic Ability Distributions

With $c = 1$, Equation 2 [with $P_i(\theta)$ from Equation 1 and $\varphi(\theta)$ from Equation 6] becomes

$$\pi_i = \int_{-\infty}^{\infty} \frac{\exp(\theta - \delta_i)\exp(\theta)}{[1 + \exp(\theta - \delta_i)][1 + \exp(\theta)]^2} d\theta. \quad (8)$$

With the substitution $t = \exp(\theta)$, the integral evaluation in Equation 8 becomes straightforward and (with simple algebra) leads to an exact formula for the expected mean on individual items:

$$\pi_i = \frac{(\delta_i - 1)\exp(\delta_i) + 1}{[\exp(\delta_i) - 1]^2}. \quad (9)$$

With $c = 1/2$, Equation 2 becomes

$$\pi_i = \int_{-\infty}^{\infty} \frac{2\exp(\theta - \delta_i)\exp(2\theta)}{[1 + \exp(\theta - \delta_i)][1 + \exp(2\theta)]^2} d\theta. \quad (10)$$

Again, using the substitution $t = \exp(\theta)$, a straightforward integration leads to an exact formula:

$$\pi_i = \frac{\pi\exp(\delta_i)[\exp(2\delta_i) - 1] - 2(2\delta_i - 1)\exp(2\delta_i) + 2}{2[1 + \exp(2\delta_i)]^2}, \quad (11)$$

where the constant π ($\approx$3.1416) is not to be confused with the notation for the domain score.

Item Error Variance with the Normal Trait Distribution

With $\varphi(\theta)$ for the standard normal *pdf*, Equation 4 can be written

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} \frac{\exp(\theta - \delta_i)}{[1 + \exp(\theta - \delta_i)]^2} \left(\frac{1}{\sqrt{2\pi}} \exp(-.5\theta^2) \right) d\theta . \qquad (12)$$

As an exact closed form evaluation for the integral in Equation 12 does not exist, an approximation was developed using the technique described with the development of Formula 7. The resulting approximation formula for the error variance of individual items is

$$\sigma^2(e_i) = A + B\exp[-0.5(\delta_i / C)^2], \qquad (13)$$

where: $A = 0.011$, $B = 0.195$, and $C = 1.797$, if $|\delta_i| < 4$,

or $A = 0.0023$, $B = 0.171$, and $C = 2.023$, if $|\delta_i| \geq 4$.

As Formula 13 shows, $\sigma^2(e_i)$ is an even function of the item difficulty, i.e., the value of $\sigma^2(e_i)$ is the same for δ_i and $-\delta_i$. Depending on the value of δ_i, the absolute error of approximation with Formula 13 ranges from 0 to 0.0008, with a mean of 0.0002 and a standard deviation of 0.0002. Also, the errors vary in sign thus canceling out to a large degree when the estimates of $\sigma^2(e_i)$ with Formula 13 are summed to obtain the error variance for the number-right score, σ_e^2 (Equation 5).

Item Error Variance with Logistic Ability Distribution

This section provides *exact* formulas for $\sigma^2(e_i)$ with the fixed logistic distributions of θ used in this article ($c = 1$ and $c = 1/2$). The mathematical derivations (provided in Appendix A) lead to the following exact evaluations of the expected item error variance, where $E_i = \exp(\delta_i)$:

1. With $c = 1$,

$$\sigma^2(e_i) = \frac{E_i(\delta_i E_i - 2E_i + \delta_i + 2)}{(E_i - 1)^3}. \qquad (14)$$

For $\delta_i = 0$, one should use $\sigma^2(e_i) = 0.1667$ (the limit evaluation with $\delta_i \to 0$) to avoid "division by zero" with Formula 14 (see Appendix A).

2. With $c = 1/2$,

$$\sigma^2(e_i) = \frac{E_i\left[8(1-\delta_i)E_i^3 + 8(\delta_i+1)E_i + \pi E_i^4 - 6\pi E_i^2 + \pi\right]}{2(E_i^2+1)^3}. \qquad (15)$$

The sum of $\sigma^2(e_i)$ for the test items is the expected error variance for the number-right score, σ_e^2.

Item True Variance

Let $\sigma^2(\tau_i)$ be the variance of the true score on item i at θ, $P_i(\theta)$, as θ varies from $-\infty$ to ∞. This *item true variance* relates to the item score, π_i, and item error variance, $\sigma^2(e_i)$, as follows

$$\sigma^2(\tau_i) = \pi_i(1 - \pi_i) - \sigma^2(e_i). \qquad (16)$$

Proof. Using the expectation rule $VAR(X) = E(X^2) - [E(X)]^2$ with $X = P_i(\theta)$, we have

$$\sigma^2(\tau_i) = \int_{-\infty}^{\infty}[P_i(\theta)]^2\varphi(\theta)d\theta - \left(\int_{-\infty}^{\infty}[P_i(\theta)]\varphi(\theta)d\theta\right)^2$$

$$= \int_{-\infty}^{\infty}\{P_i(\theta) - P_i(\theta)[1 - P_i(\theta)]\}\varphi(\theta)d\theta - \pi_i^2$$

$$= \int_{-\infty}^{\infty}P_i(\theta)\varphi(\theta)d\theta - \int_{-\infty}^{\infty}P_i(\theta)[1 - P_i(\theta)]\varphi(\theta)d\theta - \pi_i^2$$

$$= \pi_i - \sigma^2(e_i) - \pi_i^2 = \pi_i(1 - \pi_i) - \sigma^2(e_i).$$

Item Reliability

Besides reliability coefficients at test level, indices of reliability at item level can also be useful in test development and analysis. Under TSM, the reliability of item i is usually estimated with the product $s_i r_{iX}$, where s_i is the item-score standard deviation and r_{iX} is the point-biserial correlation between the item score and the total test score (e.g., Allen & Yen, 1979, p. 124). This

paper uses the definition "true item variance to observed item variance" for reliability of individual items, ρ_{ii}. Therefore, the reliability for Rasch calibrated items is evaluated here with

$$\rho_{ii} = \frac{\sigma^2(\tau_i)}{\sigma^2(\tau_i) + \sigma^2(e_i)}, \qquad (17)$$

where $\sigma^2(\tau_i)$ is obtained with Formula 16 and $\sigma^2(e_i)$ with Formula 13, when $\theta \sim N(0,1)$, or Formulas 14 and 15 when θ is with the logistic distribution for $c = 1$ and $c = 1/2$, respectively. Information about item reliability can be particularly useful when the purpose is to select items that maximize the internal consistency reliability (e.g., Allen & Yen, 1979, p. 125).

Formulas at Test Level

Marginal NR score

For a test of n binary items, the *marginal NR score*, μ, is provided with Equation 3, where the additive components (item scores, π_i) are obtained through the use of Formula 7 (the normal trait distribution), Formula 9 (the logistic trait distribution, $c_i = 1$), or Formula 10 (the logistic trait distribution, $c_i = 1/2$).

Error variance for the NR Score

The (marginal) *error variance for the NR score*, σ_e^2, is provided with Equation 5, where the additive components [item error variance, $\sigma^2(e_i)$] are obtained through the use of Formula 13 (the normal trait distribution), Formula 14 (the logistic trait distribution, $c_i = 1$) or Formula 15 (the logistic trait distribution, $c_i = 1/2$).

True Score Variance for the NR Score

The *true score variance for the NR score*, σ_τ^2, does not result from a direct summation of true variances for individual items, $\sigma^2(\tau_i)$. As proven here below, the theoretical value of σ_τ^2 is

$$\sigma_\tau^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{[\pi_i(1-\pi_i) - \sigma^2(e_i)][\pi_j(1-\pi_j) - \sigma^2(e_j)]}. \qquad (18)$$

Proof: With unidimensional tests (which are dealt with in this paper), there is a perfect correlation between the congeneric true scores on two items, say τ_i and τ_j, because of the linear relationship: $\tau_i = a_{ij} + b_{ij}\,\tau_j$, where $b_{ij} \neq 0, 1$ (e.g., Jöreskog, 1971). The covariance of τ_i and τ_j, then, is: $\sigma(\tau_i, \tau_j) = \sigma(\tau_i)\sigma(\tau_j)$. Therefore, for the variance of the true number-right score on a *n*-item test, τ ($= \Sigma\tau_i$), we have

$$\sigma_\tau^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(\tau_i,\tau_j) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(\tau_i)\sigma(\tau_j). \qquad (19)$$

Equation 19 leads directly to Formula 18 by replacing $\sigma(\tau)$ and $\sigma(\tau_j)$ with their expressions from Equation 16. It should be noted also that Formulas 16 and 18 hold for any trait distribution since their derivations remain the same with any $\varphi(\theta)$.

Reliability for Norm-Referenced Interpretations

Under TSM, the reliability of measurement is defined as the ratio of true score variance to observed score variance

$$\rho_{xx} = \sigma_\tau^2/\sigma_x^{\,2} = \sigma_\tau^2/(\sigma_\tau^2 + \sigma_\varepsilon^2). \qquad (20)$$

For internal consistency evaluations, ρ_{xx} is typically estimated by the Cronbach's coefficient alpha or by the KR20 coefficient for dichotomously scored items (Kuder & Richardson, 1937). However, even at population level, Cronbach's alpha (or KR-20) is an accurate estimate of ρ_{xx} only if there is no correlation among errors and the test components are at least essentially tau-equivalent (Novick & Lewis, 1967). For Rasch calibrated items, one can determine ρ_{xx} from Equation 20 by replacing σ_τ^2 and σ_ε^2 with their population estimates using formulas developed in the previous sections. This approach, unlike Cronbach's alpha, does not require essential tau-equivalency (the weaker assumption of congeneric measures is sufficient) thus eliminating factors that may negatively affect the population estimate of ρ_{xx} As a reminder, essentially tau-equivalent items are assumed to have equal true-score variances, whereas congeneric measures may have different scale origins and may vary in precision (Jöreskog, 1971). Previous research addresses

differences between some empirical estimates of ρ_{xx} and the *Rasch person separation reliability*, R_R (e.g., Clauser, 1999; Linacre, 1996, 1997). Both ρ_{xx} and R_R represent the ratio of "true variance to observed variance" but with ρ_{xx} the variances are for raw scores, whereas with R_R they are for trait scores (logits). Linacre (1996) reports that the true-score reliability (KR-20 or Cronbach's alpha) is generally higher than R_R, whereas the statistical Rasch validity exceeds its true-score counterpart. Also, the raw-score standard errors of extreme scores are close to zero, whereas extreme scores are usually excluded in Rasch analysis because their measure standard errors on the logit scale are infinite (e.g., Clauser, 1999).

Reliability for Criterion-Referenced Interpretations

Brennan and Kane (1977) introduced a dependability index, $\Phi(\lambda)$, for criterion-referenced interpretations in the framework of generalizability theory (GT; e.g., Brennan, 1983)

$$\Phi(\lambda) = \frac{\sigma^2(p) + (\pi - \lambda)^2}{\sigma^2(p) + (\pi - \lambda)^2 + \sigma^2(\Delta)}, \qquad (21)$$

where $\sigma^2(p)$ is the universe-score variance for persons, $\sigma^2(\Delta)$ is the absolute error variance, π is the domain score, and λ is the cutting score; (all scores are in *proportion of items correct*). In the context of the GT design "person x items", $\sigma^2(\Delta) = \sigma^2(pi,e)/n + \sigma^2(i)/n$, where n is the number of items (e.g., Shavelson & Webb, 1991, p. 86). When $\lambda = \pi$, the index $\Phi(\lambda)$ reaches its lower limit referred to as *index* Φ in GT. Feldt and Brennan (1993) noted that "the index $\Phi(\lambda)$ characterizes the dependability of decisions based on the testing procedure, whereas the index Φ characterizes the *contribution* of the testing procedure to the dependability of such decisions" (p. 141).

Taking into account that σ_τ^2 is the true variance for the person's number-right score (see Formula 18), whereas $\sigma^2(p)$ in Formula 21 is the true variance of the person's proportion of items correct, we have: $\sigma^2(p) = \sigma_\tau^2/n^2$. On the other side, $\sigma^2(i) = \sigma^2(\pi_i)$ because they both represent the variance of the expected item mean, π_i, across n items. Also, taking into account that $\sigma^2(e_i)$ is the error variance for the number-right score, the absolute error variance can be represented with

$\sigma^2(\Delta) = \sigma^2(e_i)/n^2 + \sigma^2(\pi_i)/n$. With this, Formula 21 translates into

$$\Phi(\lambda) = \frac{\sigma_\tau^2 + n^2(\pi - \lambda)^2}{\sigma_\tau^2 + n^2(\pi - \lambda)^2 + \sigma_e^2 + n\sigma^2(\pi_i)}. \quad (22)$$

When $\lambda = \pi$, $\Phi(\lambda)$ reaches its lowest limit (index Φ)

$$\Phi = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2 + n\sigma^2(\pi_i)}. \quad (23)$$

The comparison of Formulas 20 and 23 shows that Φ does not exceed $\rho_{xx'}$. This is consistent with the argument of Feldt and Brennan (1993) that "criterion-referenced interpretations of 'absolute' scores are more stringent than norm-referenced interpretations of 'relative' scores" (p. 141). It is important to emphasize that the estimation of $\rho_{xx'}$, Φ, and $\Phi(\lambda)$ in GT requires information about the raw scores for a sample of examinees, whereas the formulas developed in this article do not require such information as long as the Rasch item calibration is available.

Example

This example illustrates the estimation of expected true-score measures and reliability (at item and test level) using the formulas developed in this article for Rasch calibrated binary items. The example is organized in two sections. The first section provides (in algorithmic order) the expected measures and formulas used for their estimation with the normal trait distribution. The execution of the formulas in this section is conducted through the use of the statistical package SPSS (SPSS Inc, 1997). The SPSS syntax developed for this purpose is provided in Appendix B. The second section of this example compares the expected true-score measures and reliability to their empirical counterparts obtained with simulated data.

Theoretical Evaluation of True-Score Measures with Formulas

This section illustrates how researchers and practitioners may use the Rasch calibration of binary items to evaluate expected true-score measures and reliability at both item and test level.

The Rasch difficulty parameters, δ_i, for 20 hypothetical items are provided in Table 1; (δ_i sum to zero and cover uniformly the interval from -2.2 to 2.5 on the logit scale). The expected measures and the formulas used for their evaluation with $\theta \sim N(0,1)$ are listed below in algorithmic order.

1. *Item mean*, π_i - Formula 7.
2. *Item error variance*, $\sigma^2(e_i)$ - Formula 13.
3. *Item true variance*, $\sigma^2(\tau_i)$ - Formula 16.
4. *Item reliability*, ρ_{ii} - Formula 17.
5. *Marginal number-right score*, μ - Formula 3; (the *domain score* is $\pi = \mu/n$).
6. *Error variance for the number-right score*, σ_e^2 - Formula 5.
7. *True score variance for the number-right score*, σ_τ^2 - Formula 18.
8. Descriptive variance of the item scores, $\sigma^2(\pi_i)$ - the variance of $\pi_1, \ldots, \pi_{20}$ (see Step 1).
9. *Reliability*, ρ_{xx} - Formula 20.
10. *Dependability index* Φ - Formula 23.
11. *Dependability index* $\Phi(\lambda)$ - Formula 22.

The SPSS printout (with the syntax in Appendix B and item parameters, δ_i, in Table 1) provides the expected true variance for the number-right score ($\sigma_\tau^2 = 10.4200$), the error variance for the number-right score ($\sigma_e^2 = 3.1046$), the expected number-right score ($\mu = 10.1027$), and the variance of expected item means, $\sigma^2(\pi_i) = .071$. Using these values, we obtain: $\pi = \mu/n = .5051$, $\rho_{xx} = .7704$ (with Formula 20), and $\Phi = .6972$ (with Formula 23). Also, using Formula 22, values of the dependability index $\Phi(\lambda)$ are calculated and graphed for values of the cutting score, λ, that vary from 0 to 1 on the domain scale with an increment of 0.005 (see Figure 2). The graphical representation of $\Phi(\lambda)$ shows, for example, that its lowest value ($\Phi = .6972$) occurs when the cutting score equals the population domain score ($\lambda = \pi = .5051$). Also, $\Phi(\lambda) = .85$ for $\lambda = .7$ and $\Phi(\lambda)$ exceeds .90 when the cutting score is above .8 (i.e., 80% in percentages). This type of information is very useful for criterion-based interpretations and decisions with mastery tests.

The SPSS syntax (see Appendix B) provides also the expected true-score measures and reliability for individual items. They appear as "new" variables in the SPSS data spreadsheet, with

notations that should be interpreted as follows: var_e = $\sigma^2(e_i)$, p = π_i, var_tau = $\sigma^2(\tau_i)$, and roi = $\rho_{ii'}$ (the values of these output variables are provided in Table 1).

Empirical Validation of the Formulas with Simulated Data

The expected measures obtained in the previous section are compared now with their empirical counterparts obtained with simulated data. Specifically, binary scores were generated to fit the Rasch model [with the item parameters, δ_i, in Table 1 and $\theta \sim N(0,1)$] using a computer program written in SAS (SAS Institute, 1985) for Monte Carlo simulations (Dimitrov, 1996). The (ANOVA-based) generalizability model "person x item" ($p \times i$) incorporated in this program was run with 20 replications generating binary scores for 1,5000 persons in each replication. The resulting empirical estimates of true-score measures and reliability (at test level) are summarized in Table 2. The comparison of these empirical estimates with their theoretical counterparts (also presented in Table 2) shows a close match. The same holds for the comparison of the expected item means, π_i, with their empirical counterparts (p_i) obtained for the SAS simulated binary scores (see Table 1). Thus, with Rasch calibrated items, the formulas developed in this article provide (without data) estimates of true-score measures and reliability that one can obtain (with "ideal" data simulated for large samples) using the "person x item" GT model. In addition, the formulas provide the marginal values of true-score measures and reliability for individual items [$\sigma^2(\tau_i)$, $\sigma^2(e_i)$, and ρ_{ii}] that are not provided with the GT model.

The Rasch person separation reliability index , R_R, was also calculated for the generating measures and item difficulties with the SAS simulations. Linacre (1997) refers to R_R obtained with generated θ- measures as *generator-based Rasch reliability* and shows that it is an upper limit for data-based R_R. The generator-based reliability with the SAS simulations in this example was found to be R_R = .673. The fact that the theoretical ρ_{xx} (.770) is higher than R_R (.673) in this example is not a surprise given that even empirical estimates of ρ_{xx} (KR-20 or Cronbach's alpha) generally exceed R_R (Linacre, 1996).

Table 1

True-Score Measures and Reliability for Individual Items Evaluated as a Function of Their Rasch Difficulty, δ_i.

Item	δ_i	$\sigma^2(e_i)$	π_i	$(p_i)^a$	$\sigma^2(\tau_i)$	ρ_{ii}
1	-2.2000	.1032	.8656	(.8620)	.0132	.1131
2	-2.0000	.1160	.8440	(.8446)	.0157	.1191
3	-1.8200	.1278	.8224	(.8133)	.0183	.1251
4	-1.5300	.1467	.7833	(.7920)	.0231	.1358
5	-1.4000	.1550	.7639	(.7560)	.0254	.1408
6	-1.2500	.1641	.7402	(.7287)	.0282	.1466
7	-1.0500	.1754	.7065	(.6787)	.0320	.1541
8	-.8500	.1854	.6705	(.6827)	.0356	.1610
9	-.6100	.1951	.6247	(.6054)	.0394	.1679
10	-.2500	.2041	.5520	(.5600)	.0432	.1746
11	.0000	.2060	.5000	(.4874)	.0440	.1760
12	.2800	.2036	.4419	(.4463)	.0430	.1742
13	.4500	.2000	.4072	(.3860)	.0414	.1715
14	.8500	.1854	.3295	(.3400)	.0356	.1610
15	1.2100	.1664	.2663	(.2514)	.0289	.1481
16	1.3300	.1593	.2470	(.2353)	.0267	.1435
17	1.9700	.1179	.1594	(.1633)	.0161	.1201
18	2.1500	.1063	.1395	(.1300)	.0138	.1145
19	2.2200	.1019	.1323	(.1180)	.0129	.1125
20	2.5000	.0851	.1064	(.1093)	.0100	.1049

Note. $\sigma^2(e_i)$ is the expected error variance, π_i - expected mean, (p_i - empirical mean), $\sigma^2(\tau_i)$ - expected true variance, and ρ_{ii} - expected reliability for individual items.

[a] Obtained for the SAS simulated binary scores.

Table 2

Theoretical True-Score Measures and Reliability (Evaluated with Formulas) and Their Empirical Counterparts Evaluated with Simulated Data for the Rasch Item Difficulties (δ_i) in Table 1 and the Normal Trait Distribution.

Evaluation	π	σ_t^2	σ_e^2	$\sigma^2(\pi_i)$	ρ_{xx}	Φ
Theoretical	.5051	10.4200	3.1046	.0710	.7704	.6972
Empirical	.5067	9.9572	3.1647	.0708	.7548	.6802

Note. The empirical estimates are obtained through averaging their values over 20 replications of SAS simulations for binary scores that fit the Rasch model with 1,500 persons per replication.

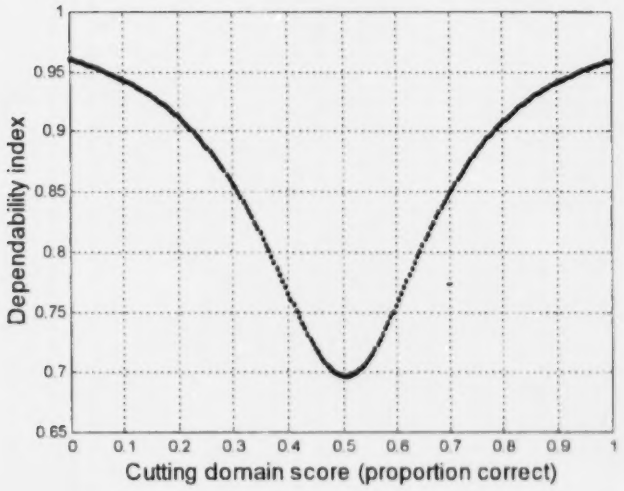

Figure 2. The dependability index, $\Phi(\lambda)$, estimated with Formula 21 for the theoretical true-score measures in Table 2 with the illustrative example.

Conclusion

This paper provides formulas for true-score measures and reliability of binary scores as a function of the Rasch item difficulty for fixed distributions (normal or logistic) of the underlying trait. The scale parameters $c = 1$ and $c = 1/2$ were selected for the two fixed logistic distributions because they yield exact integral evaluations and produce normal-like trait distributions of the underlying trait that may occur with Rasch measurements; (this is not true with just any scale parameter of the logistic distribution). Formulas 7 and 13 for π_i and $\sigma^2(e_i)$, respectively, with the normal trait distribution are developed by the use of approximation procedures, whereas all other formulas result from exact integral evaluations. The example in the previous section illustrates an application of the formulas for Rasch calibrated items. The calculations are easy to perform using statistical programs such as SAS and SPSS (see Appendix B), spreadsheet-based programs, or even hand calculators. The formulas can also be efficiently incorporated into computer programs for test analysis and measurement simulations.

The formulas developed in this paper have theoretical and practical value for Rasch test development, score analysis, and simulation studies. Their closed analytical forms may reveal relationships that are difficult or impossible to see with empirical tools (e.g., Formula 13 shows that the item error variance has the same value for opposite, δ_i and $-\delta_i$, Rasch item difficulties). Also, given a bank of Rasch calibrated items, one can select items to develop a test with known true-score measures and reliability for a person population *prior* to administering the test. One can also compare (without using raw scores or trait measures) the expected domain scores and reliability for test strands in which items are grouped by substantive characteristics (e.g., content areas or learning outcomes). In another scenario, the formulas can be used to evaluate (*prior* to administration) test booklets that are developed for follow-up measurements (e.g., in longitudinal studies) given the Rasch calibration of items at the base year. In simulation studies, researchers may use the formulas to generate true-score characteristics and reliability for targeted values of Rasch item difficulty without the necessity of generating binary scores or θ-scores for persons.

The examples of possible applications of the formulas developed in this paper illustrate

what researchers and practitioners can gain over and above what they would learn from the Rasch analysis. It is important to emphasize that the proposed formulas and the Rasch analysis provide different types of information that can efficiently complement (not replace or exclude) each other in test development and analysis. For example, while the Rasch analysis is effective at locating persons on the underlying trait (Linacre, 1996), the formulas developed in this article are effective at determining population true-score characteristics for Rasch calibrated items without using raw scores or trait measures for examinees. Also, while the Rasch measures of reliability (R_R) and "separation" provide information about measurably different levels of performance in a sample of examinees (e.g., Wright, 1996, 1998), the index $\Phi(\lambda)$ provides information about the dependability of criterion-referenced decisions. Which approach to use (Rasch analysis, true-score analysis with the proposed formulas, or both) depends on the goals of the study as well as on the data that is available (raw scores, trait scores, or only estimates of Rasch item difficulty).

One can also argue that estimates of true-score measures and reliability can be obtained within the framework of generalizability theory using, for example, computer programs such as GENOVA (Crick & Brennan, 1983). This approach, however, (a) requires the binary scores for a large sample of examinees and (b) does not provide true-score measures at item level such as $\sigma^2(e_i)$, $\sigma^2(\tau_i)$, and $\rho_{ii'}$. Therefore, for Rasch calibrated items, the formulas developed in this paper provide (without data) richer, more accurate, and easily obtained information about true-score measures and reliability at population level relative to (ANOVA-based) generalizability methods.

Skewed trait distributions also occur with Rasch measurement (e.g., in medical studies; Wright, 2001). Dimitrov (2001) provided formulas for the expected error variance with some skewed trait distributions. Formulas 16 and 18 for the true score variance can also be used with skewed distributions because their derivation holds with any $\varphi(\theta)$. In conclusion, using Rasch calibration of items to evaluate their expected true-score measures, reliability, and dependability extends the traditional boundaries in calculating, interpreting, and reporting measurement results.

References

Allen, J. M., & Yen, W. M. (1979). *Introduction to measurement theory*. Pacific Grove, CA: Brooks/Cole.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Brennan, R. L., & Kane,M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277-289.

Clauser, B. (1999). Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions, 13:2*, 696.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system*. The American College Testing Program, Iowa City: Iowa.

Cronbach, L. J. (1951). Coeficient alpha and the internal structure of a test. *Psychometrika, 16*, 297-334.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*, 783-801.

Dimitrov, D. M. (2001, October). Reliability of Rasch measurement with skewed ability distributions. Paper presented at the International Conference on Objective Measurement. Chicago, IL.

Dimitrov, D.M. (1996, April). *Monte Carlo approach for reliability estimations in generalizability studies*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Evans, M.., Hastings, N., and Peacock, B. (1993). *Statistical distributions* (2nd ed.). NY: John Wiley.

Feldt, L. S., and Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd. Ed.) Phoenix, AZ: American Council on Education and The Oryx Press (pp. 105-146).

Hambleton, R. K., and Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 1937, 2*, 151-160.

Linacre, J. M. (1997). KR-20 or Rasch reliability: Which tells the "truth"? *Rasch Measurement Transactions, 11:3*, 580-581.

Linacre, J. M. (1996). True-score reliability or Rasch statistical validity? *Rasch Measurement Transactions, 9:4*, 455-456..

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NL: Lawrence Erlbaum.

MathWorks, Inc. (1999). *Learning MATLAB (Version 5.3)*. Natick, MA: Author.

Novick, M. R., and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.

Rasch, G. (1960). *Probabilistic models for intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

SAS Institute. (1985). *SAS user's guide: Version 5 edition*. Cary, NC: Author.

Sawilowski, S. S. (2000). Psychometrics versus datametrics: Comments on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement, 60*, 157-173.

Shavelson, J. S., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Smith, Jr., E. V. (2000). Metric development and score reporting. *Journal of Applied Measurement, 3*, 303-326.

Smith, Jr., E. V. (2001). Evidence for the reliability of measures and validity of measure

interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281-311.

SPSS Inc. (2002). *SPSS Base 11.0 User's guide*. Chicago: Author.

SPSS Inc. (1998). *SigmaPlot 5.0 User's guide*. Chicago: Author.

Thompson, B., and Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

Wright, B. D. (2001). Separation, reliability and skewed distributions. *Rasch Measurement Transactions 14:4*, 786.

Wright, B. D. (1998). Interpreting reliability. *Rasch Measurement Transactions, 11:4*, 602.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions, 9:4*, 472.

Appendix A

Derivation of Formulas 14 and 15 for Item True Variance with Logistic Trait Distribution

For $P_i(\theta)$ with the dichotomous Rasch model (Equation 1), we have

$$P_i(\theta)[1 - P_i(\theta)] = \frac{\exp(\theta - \delta_i)}{[1 + \exp(\theta - \delta_i)]^2} \qquad \text{(A1)}$$

which (as one can easily see) is also the first derivative of $P_i(\theta)$. With this, Equation 4 becomes

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} [\partial P_i(\theta) / \partial\theta]\varphi(\theta)d\theta = \int_{-\infty}^{\infty} \varphi(\theta)dP_i(\theta). \qquad \text{(A2)}$$

As one may also notice, the logistic $\varphi(\theta)$ in Equation 6 is the first derivative of the function

$$\Phi(\theta) = \frac{\exp(\theta / c)}{1 + \exp(\theta / c)}.$$

Replacing $\varphi(\theta)$ in Equation A2 with the first derivative of $\Phi(\theta)$, we have

$$\sigma^2(e_i) = \int_{-\infty}^{\infty} [\partial P_i(\theta) / \partial\theta][\partial\Phi(\theta) / \partial\theta]d\theta = \int_{-\infty}^{\infty} [\partial P_i(\theta) / \partial\theta]d\Phi(\theta). \qquad \text{(A3)}$$

With integration by parts for the integral in Equation A3, we have

$$\sigma^2_{e_i} = [\partial P_i(\theta) / \partial\theta]\Phi(\theta)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \Phi(\theta)d[\partial P_i(\theta) / \partial\theta]$$

$$= 0 - \int_{-\infty}^{\infty} \Phi(\theta)[\partial^2 P_i(\theta) / \partial\theta^2]d\theta$$

$$= -\int_{-\infty}^{\infty} \frac{\exp(\theta / c)\exp(\theta - \delta_i)[1 - \exp(\theta - \delta_i)]}{[1 + \exp(\theta / c)][1 + \exp(\theta - \delta_i)]^3} d\theta.$$

Let $E_i = \exp(\delta_i)$. Using the substitution rule for integration with $x = \exp(\theta)$, we obtain

$$\sigma^2(e_i) = \int_0^{\infty} \frac{E_i x^{1/c}(x - E_i)}{(1 + x^{1/c})(x + E_i)^3} dx. \qquad (A4)$$

The evaluation of the integral in Equation A4 for $c = 1$ or $c = 1/2$ is straightforward and yields to

1. With $c = 1$,

$$\sigma^2(e_i) = \frac{E_i(\delta_i E_i - 2E_i + \delta_i + 2)}{(E_i - 1)^3}. \qquad (A5)$$

When $\delta_i = 0$, the denominator of the ratio in Formula A5 equals zero. For this particular case, estimating the limit of the ratio at $\delta_i = 0$, we obtain $\sigma^2(e_i) = 0.1667$.

2. With $c = 1/2$,

$$\sigma^2(e_i) = \frac{E_i\left[8(1 - \delta_i)E_i^3 + 8(\delta_i + 1)E_i + \pi E_i^4 - 6\pi E_i^2 + \pi\right]}{2(E_i^2 + 1)^3}, \qquad (A6)$$

where π is a constant ($\pi = 3.14159...$ is not to be confused with the domain score) and E_i denotes $\exp(\delta_i)$ for simplicity of the analytical form. As one may notice, Formulas A5 and A6 are exactly Formulas 14 and 15, respectively, with which the derivation is completed.

Appendix B

SPSS Syntax for Evaluation of True-Score Measures of Rasch Calibrated Binary Items with the Normal Trait Distribution; (Input variable: **b,** the Rasch item difficulty)

```
DO IF (ABS(b) < 4).
COMPUTE ve = .011 + .195*exp(-.5*((b/1.797)**2)).
ELSE.
COMPUTE ve = .0023 + .171*exp(-.5*((b/2.023)**2)).
END IF.
COMPUTE p = -.0114 + 1.0228/(1 + exp(b/1.226)).
COMPUTE vt = p*(1 - p) - ve.
IF(vt < 0) vt = 0.
SET FORMAT = F8.4 ERRORS = NONE RESULTS OFF HEATHER NO.
FLIP
  VARIABLES  b ve p vt.
VECTOR V = VAR001 TO VAR020.
COMPUTE Y = 0.
LOOP  #I = 1 TO 20.
LOOP #J = 1 TO 20.
     COMPUTE Y = Y + SQRT(V(#I)*V(#J)).
END LOOP.
END LOOP.
FLIP VAR001 TO VAR020 Y.
COMPUTE roi = vt/(vt + ve).
SET RESULTS ON.
REPORT FORMAT = AUTOMATIC
   /VARIABLES = ve ' ' p ' ' vt ' '
   /BREAK = (TOTAL)
   /SUMMARY = MAX(vt) 'True score variance:'
   /SUMMARY = SUBTRACT(SUM(ve) MAX(ve)) (vt (COMMA) (4)) 'Error variance:'
   /SUMMARY = SUBTRACT(SUM(p) MAX(p)) (vt (COMMA) (4)) 'Expected mean:' .
SELECT IF(CASE_LBL ~= 'Y' ) .
RENAME VARIABLES (CASE_LBL = ITEM) (ve = var_err) (vt = var_tau).
VARIABLE LABELS p 'Expected item mean' .
DESCRIPTIVES
    VARIABLES = p
    /STATISTICS = VAR .
```

Note. The number of items (in this example, 20) should be specified in the syntax by the user. With 50 items, for example, change **20** to 50 and **VAR020** to VAR050 (see the bold notations in the respective four syntax lines).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034865

I. DOCUMENT IDENTIFICATION:

Title: Reliability of True Cutting Scores for Rasch Calibrated Items	
Author(s): Dimiter M. Dimitrov	
Corporate Source:	Publication Date: April 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

[X]

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

[ ]

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

[ ]

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Dimitrov	Printed Name/Position/Title: Dimiter Dimitrov / Assoc. Professor
Organization/Address: 507 Whitehall, College of Ed, Kent, OH 44242-0001	Telephone: 330-672-0582 FAX:
	E-Mail Address: DDIMITRO@Kent.edu Date: 4/2003

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse: **University of Maryland**
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions